OCAIML4001: Responsible & Safe AI Systems

Description:

This course OCAIML4001: Responsible & Safe AI Systems is offered from SWAYAM

URL : https://onlinecourses.nptel.ac.in/noc24_cs132/preview

Credits and Hours:

Teaching Scheme	Theory	Practical	Total	Credit
Hours/week	3	2	5	**4**
Marks	100	50	150	

*Practical component is offered by CHARUSAT

About this course:

This course provides students with a comprehensive understanding of the ethical, social, and safety considerations essential for developing and deploying artificial intelligence (AI) systems. Uncover the intricacies of algorithmic transparency, fairness in machine (un)learning, interpretability, consistency and many more. The course encourages critical thinking and fosters a deep appreciation for the impact of AI on individuals and communities. Students who complete the course can: recognize possible harms that can be caused by modern AI capabilities; learn to reason about various perspectives on the trajectory of AI development and proliferation; learn about latest research agendas towards making AI systems safer.

Course Layout:

Week 1 & 2: AI Capabilities Improvement in last 5-10 years , Imminent risks from AI Models: Toxicity, bias, goal misspecification, adversarial examples etc. , Long-term risks from AI Models: Misuse, Misgeneralization, Rogue AGI , Principles of RAI - Transparency; Accountability; Safety, Robustness and Reliability;Privacy and Security; Fairness and non-discrimination; Human-Centred Values; Inclusive and Sustainable development, Interpretability, Recap of Deep Learning Techniques, Language/Vision Models **,** AI Risks for Gen models, Adversarial Attacks – Vision, NLP, Superhuman Go agents.

Week 3 & 4: ML Poisoning Attacks like Trojans ,Implications for current and future AI safety , Explainability ,Imminent and Long-term potential for transparency techniques, Mechanistic Interpretability , Representation Engineering, model editing and probing, Critiques of Transparency for AI Safety

Week 5 & 6: Privacy & Fairness in AI

Week 7 & 8: Metrics and Tools for RAI - measuring bias/fairness, adversarial testing, explanations (Lime/SHAP/GradCam), audit mechanisms ,Regulation landscape - DPDP act (India), GDPR (EU), EU AI act, US presidential declaration, Ethical approvals, informed consent, participatory design, future of work, Indian context ,What is AGI? When could it be achieved? ,Instrumental Convergence: Power Seeking, Deception etc.

Week 9 & 10: RAI in Legal domain , RAI in Health care domain , RAI in Education domain , A few other domains ,Policy issues in RAI .

Week 11 & 12: Couple of panel discussion with industry practitioners, academic, government (possibly), and others , Fireside chat with eminent personalities , Recorded Paper reading discussion

Course Outcomes (COs):

CO1	Identify ethical, legal, and social issues related to the development and deployment of AI systems.
CO2	Analyze real-world AI applications to assess fairness, transparency, and potential biases.
CO3	Demonstrate understanding of algorithmic accountability, interpretability, and decision-making consistency.
CO4	Evaluate the safety and robustness of AI systems using contemporary tools, frameworks, and case studies.
CO5	Apply principles of responsible AI in the design and critique of machine learning models and data pipelines.
CO6	Interpret and discuss current research and policy directions in AI safety, governance, and regulation.
CO7	Communicate informed perspectives on the ethical implications and future trajectory of AI technologies.

Course Articulation Matrix:

Course Outcomes (COs)	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	2	1	–	–	–	3	3	3	–	2	–	2
CO2	3	2	2	–	–	3	3	3	1	2	–	2
CO3	3	2	2	–	–	3	2	2	1	2	–	2
CO4	3	3	3	2	2	2	2	3	2	2	–	2
CO5	3	2	3	2	2	3	3	3	1	3	1	3
CO6	2	2	1	–	–	3	3	3	2	2	–	2
CO7	2	1	–	–	–	3	3	3	3	3	–	2